CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Springer - Publisher Connector

Cogn Tech Work (2015) 17:503–519
DOI 10.1007/s10111-015-0325-3

CrossMark

ORIGINAL ARTICLE

Modelling environmental and cognitive factors to predict performance in a stressful training scenario on a naval ship simulator

Iris Cohen · Willem-Paul Brinkman · Mark A. Neerincx

Received: 7 March 2014 / Accepted: 23 January 2015 / Published online: 12 February 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract Professionals working in risky or emergency situations have to make very accurate decisions, while the quality of the decisions might be affected by the stress that these situations bring about. Integrating task feedback and biofeedback into computer-based training environments could improve trainees' stress-coping behaviour. This paper presents and assesses a refined version of the cognitive performance and error (COPE) model that describes the effects of stressful events on decisions as a foundation for such a support tool. Within a high-fidelity simulator of a ship's bridge at the Royal Netherlands Naval College, students of the naval college ($n = 10$) were observed while completing a 2-h-long shadowing and boarding operation combined with a search-and-rescue operation. For every action, variables were measured: objective and subjective task demand, challenge and threat appraisal, and arousal based on heart rate and heart rate variability. The data supported the COPE model and were used to create predictive models. The variables could provide minute-by-minute predictions of performance that can be divided into performance rated by experts and errors. The predictions for performance rated by experts correlated with the observed data ($r = 0.77$), and 68.3 % of the predicted errors were correct. The error predictions concern the chances of making specific errors of *communication*, *planning*, *speed*, and *task allocation*. These models will be implemented into a real-time feedback system for trainees performing in stressful simulated training tasks.

Keywords Stress · Virtual training · Cognitive errors · Performance · Simulator · Navy

I. Cohen (✉) · M. A. Neerincx
TNO, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
e-mail: iris.cohen@tno.nl

I. Cohen · W.-P. Brinkman · M. A. Neerincx
Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

1 Introduction

Professionals working in safety-related fields such as the police force, fire department, aviation, and the army may enter uncertain and unexpected situations that bring along high levels of stress and demands (Driskell and Johnston 1998). For example, naval ship operators encounter situations where they have to process a great amount of complex information in a short period of time and make a decision that can have severe consequences. Unfortunately, high levels of stress can negatively affect cognitive functions that are needed to execute several cognitive processes (Mendl 1999). For example, errors are likely to occur in cognitive functions such as: attention, memory formation, and memory recall (Kleider et al. 2010; Mendl 1999; Orasanu and Backer 1996). In order to mitigate negative effects of stress, it is important to understand (1) the underlying processes and their effects on performance and (2) the experiences with decision support systems that have been developed to improve performance. Understanding these two topics will help to achieve the aim of this paper: establishing predictive models that can be used in a new decision or training support system. This introduction starts with an overview of the literature on decision-making under stress. Next, past and current decision support systems and other training methods are discussed to give an idea on what is important when designing such a system. The introduction then ends with

a more detailed formulation of the research aim and hypotheses of this paper.

Decision-making involves a specific *cognitive process* that is influenced by high stress levels (Starcke and Brand 2012; Kerstholt 1994). Considering alternative decision options is a step in the decision-making process where stress can have negative effects. Individuals are more likely to decide without considering all alternatives (*premature closure*), use a non-systematic manner to consider the alternatives (*non-systematic scanning*), and seem unable to allocate time to consider all the alternatives (*temporal narrowing*) (Keinan et al. 1987). Time constrains seem to play a key role in these circumstances. For example, Maule et al. (2000) found that time pressure induced feelings of being energetic and anxious in people. But time pressure is not a prerequisite for stress. Keinan et al. (1987) reported that people can show disorganized and incomplete scanning when time limits are not present. Another observation relevant to these situations is that making a decision should not be seen as a single action, but as a chain of unfolding events and decisions. Ozel (2001) mentioned that human behaviour seems to be episodic in stressful and dangerous events. Every episode focuses on a certain goal that needs to be reached by executing appropriate actions. Achieving the goals can be seen as 'decision-making between episodes', and achieving the actions can be seen as 'decision-making within episodes'. Distinguishing goals and actions in human behaviours during emergency handling makes it easier to investigate where in the decision processes stress plays a role. Another aspect of professionals working in stressful environments is that professionals often operate in teams. Working in a team can have obvious benefits, but also brings along extra-cognitive issues that can have negative effects on performance during team decision-making. Dowell and Hoc (1995) group these cognitive issues of coordinating decision-making and actions in four groups: planning, action, communication, and task knowledge.

Current practices aiming to reduce negative effects of stress make use of technical advances such as *decision support systems* or training environments that induce stress. Since the early 80s, research has tried to create effective digital decision support systems, or Intelligent Decision Aids (IDAs) (Kontogiannis and Kossiavelou 1999). Early support systems were designed to create decisions without biases. These systems provided limited options for the users to assess system's outcome: the users could merely accept or reject the decision made for them. This might have been a reason that the users had problems accepting these kinds of decisions and support systems (Kontogiannis and Kossiavelou 1999). Other problems were that the decision tools, even when focussed on naturalistic decisions, rarely showed decision improvement because individuals using them were often ahead of the tool (Cohen 1993), and the tool designers cannot anticipate all possible scenarios that might occur (Reason 1987). Therefore, recent and current IDAs are being designed to collaborate with its users to reach decisions, e.g. aiming at a 'joint (human technology) cognitive system' (cf. Hollnagel and Woods 2005). In their review, Kontogiannis and Kossiavelou (1999) also propose that IDAs should try to prevent and delay stress. This can be done by implementing suggestions for changes in team strategies proven to be efficient while working under stress into IDAs. IDAs should provide insight into event escalations and the anticipation of rare events. They should point out changes in communication necessary to work under stress and help the team members to keep track of each other's activities. Also the structure and task allocation of teams should adapt to stressful situations.

Another approach to prepare professionals to stressful environments is to expose them to stressful conditions during *scenario-based training*, so that they can learn to cope with such conditions and to keep their performances at a high level in a stressful environment (Driskell and Johnston 2006; Peeters et al. 2014). Previous research has found several aspects that can be applied to create effective stress training. First, training environments should clearly convey a naturalistic environment. Making decisions in a real-life event is hardly the same as making decisions in a laboratory setting on which the classical decision theory is based (Beach and Lipshitz 1993). Orasanu and Connolly (1993) listed eight factors that have been ignored in decision research, but are clear features of decision-making in a naturalistic environment. The factors they list are as follows: ill-structured problems, uncertain dynamic environments, shifting or competing goals, action or feedback loops, time stress, high stakes, multiple players and organizational goals, and norms. The presence of several of these factors in stressful situations will complicate the task of making a decision. Besides properties of naturalistic environments, specific guidelines have been suggested with regard to simulation training. For example, Sime (2007) listed seven properties for simulation training that help to reduce stress and its negative effects on decisions. Her seven suggestions are as follows: (1) when training certain skills that are to be applied in a stressful environment, the training setting should be a stressful environment as well; (2) reducing workload caused by time pressure can be achieved by rehearsing cognitive and behavioural skills up to automation; (3) by training heuristics of task prioritization; (4) cognitive rehearsal of a task can help increase one's confidence and ability; (5) while team training increases team performance through the sense of team identity; (6) changing the training environments helps train flexibility, which makes it easier to work in an unknown

situation; and last (7) negative emotions and fear of the unknown can be reduced with the right training such as biofeedback training and cognitive control strategies.

Besides training in naturalistic environments, Sime (2007) suggested that *biofeedback* can be an effective tool to decrease stress during training. Whereas biofeedback increases control over one's physiological stress reactions (Bouchard et al. 2012), e.g. increased heart rate and fast respiration, cognitive control strategies can reduce emotions and distracting thoughts (Sime 2007; Driskell and Johnston 2006). Having a clear understanding of one's emotions will help individuals to experience fewer cognitive difficulties. It is argued that when under stress, cognitive attention resources will not only be depleted by the task at hand, but also be depleted by the emotional reactions (Gohm et al. 2001; Driskell and Johnston 2006). When less cognitive resources are available, performance will decline. In other words, a better insight into one's emotional reactions improves performance under stress.

The project 'better decisions under high pressure' was started to develop computer-based training support for mitigating negative effects of stress on decision-making. The envisioned support tool incorporates above-mentioned training and biofeedback approaches, i.e. by *combining* biofeedback (Sime 2007), and suggestions for changes in strategies (Kontogiannis and Kossiavelou 1999) and cognitive control strategies. Using only biofeedback teaches individuals to control their physiological reactions to stress, but not their cognitive reactions (Mendl 1999; Keinan et al. 1987; Gohm et al. 2001). Cognitive feedback by suggesting efficient team strategies, together with biofeedback, could help trainees to overcome cognitive issues or impairments due to stress. In addition, it is expected that a tool that provides such combined support will be accepted better by the end-users.

To establish the real-time biofeedback and performance feedback, a model is needed that assesses the task and emotional load and provides performance predictions. The first model development step is to combine situational factors and cognitive and physiological indicators in a descriptive model and, subsequently, to refine it into a predictive model for cognitive processes and performances that are likely to occur in certain stressful situations. Cohen et al. (2012) provided a first (descriptive) version of this model based on the literature on cognitive reactions to stress, called the COgnitive Performance and Error (COPE) model. The goal of this study is to validate a refined version of the COPE model and test its ability to predict cognitive errors and performance. This paper describes the acquisition of training data and the subsequent analysis of the relationships between the COPE variables. The first hypothesis states that the variables are related as suggested by the COPE model. The second hypothesis states that the cognitive and situational variables in the COPE model can be used to predict performance and errors under stress. The next section of this paper will describe the variables of the COPE model and their expected relations.

2 COPE model

The graphical representation of the COPE model displayed in Fig. 1 shows a cognitive process of decision-making under stress (Cohen et al. 2012). It roughly consists of three components: the external world, the individual's cognition and the individual's actions interpreted as the performance on a task or decision.

In this model, the external environment consists of an *event* and the corresponding *objective task demand*. An event itself is not stressful, but an individual can experience an event as a stressful event. Whether an event is experienced as a stressful one or not depends on the individual's cognitive perception of the event. The task demand variables are based on Neerincx's (2003) model of Cognitive Task Load. In this model, task demand is divided into three dimensions: level of information processing, time occupied, and task-set switches. By measuring these three dimensions, it is possible to determine cognitive task load during a specific task. The distinction between *objective* and *subjective task demand* implies that task demands can be determined 'from the outside', e.g. by external experts or task analysts (called 'objective') and by the task performers themselves (called 'subjective'). The subjective task demands can be lower or higher than the objective task demands (Bosse et al. 2008).

Stress reactions that follow a stressful event can be explained as indirect reactions to the stressful event (Lazarus 1999). After perceiving a stressful event, the severity of potential danger is assessed by the person experiencing it. This assessment is called the *primary appraisal*. If a situation is appraised as dangerous, it can be seen as a *challenge* when the individual feels he or she can cope with the event, or as a *threat* when the individual feels he or she is lacking the resources to cope with the event. This is called the *secondary appraisal*. An individual that is experiencing a situation appraised as a threat or a challenge will try to cope with the situation by applying an appropriate *coping strategy* (Gaillard 2007). Which coping strategy is used by the individual depends on the appraisal, but also on the individual's emotional state, since affect influences judgment (Forgas 1995). The chosen coping strategy, in its turn, influences the decisions and actions made by the individual (Delahaij 2009). Thunholm (2004, 2008) investigated individual's decision-making styles while under stress and found that an avoidant decision style relates to higher levels of distress

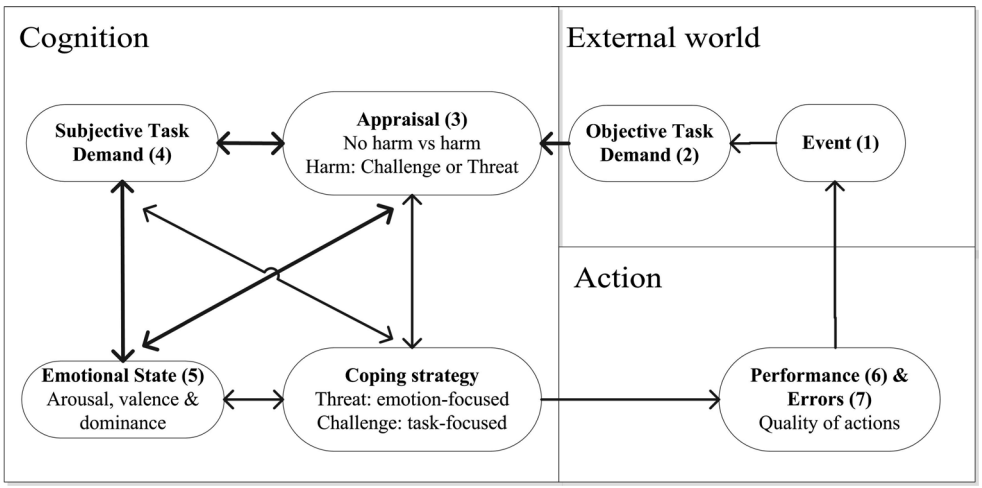

Fig. 1 Schematic view of the COPE model of external and cognitive factors, predicting an individual's performance and errors

and that a spontaneous decision style did not. Although decision-making styles and coping strategies fit in the COPE model, they are out of the scope of this study, since there are no quick and easy ways to determine which style is used by the trainees.

A common way of measuring *Emotional State* is by using the *valence, arousal*, and *dominance* scale (Bradley and Lang 1994; Mehrabian 1996). While valence is a scale that indicates the pleasantness of stimuli experienced by an individual, the arousal scale ranges from being excited to relaxed. The dominance scale represents the level of control an individual feels. Instead of using a questionnaire, arousal can be measured in a less obtrusive way by measuring physiological aspects using biosensors (Haag et al. 2004). Physiological measures related to arousal induced by stress are, for example, heart rate (HR), heart rate variability (HRV), and stress hormone levels (Krantz et al. 2004; Hjortskov et al. 2004).

HR increases due to the sympathetic nervous system (SNS) stimulation caused, for example, by stress, exercise, or cardiovascular disease. Activation of the parasympathetic nervous system (PNS) causes a decrease in HR. Changes in the balance between PNS and SNS activation produce heart rate fluctuations known as HRV. HR and HRV are used in the literature as measures of mental effort; an increase in mental effort will increase HR and decrease HRV (Mulder 1992). Mulder (1992) described a decrease in HRV as invested effort and not just a higher task difficulty. The effort needed to perform a more difficult task is shown by lowered HRV.

At the end of the cycle, an individual's cognition will lead to certain decisions and actions. Whether these decisions or actions are appropriate for the stressful event will determine the performance on the task. Reacting to the event will eventually result in changes of the external world and new tasks to perform and decision to make.

3 Methods

After the explanation of the COPE model in the previous sections, the hypotheses can be described in more detail. The first hypothesis states that the arrows in Fig. 1 represent correlations between the variables. The second hypotheses states that the cognitive variables (appraisal, task demand, and physiological arousal) and the objective task demand can predict performance values.

To validate the COPE model and use the variables to predict performance and cognitive errors under stress, seven variables from the COPE model were measured (Sect. 3.3) while participants performed tasks in a stressful virtual scenario. The scenario took place in two simulated ship environments at the Royal Netherlands Naval College (RNNC) in Den Helder, The Netherlands. In every session, two teams of three participants were formed, each team in a separate simulator (simulators were connected). They experienced the same stressful scenario in which they needed to make decisions and execute tasks that would lead to a positive outcome.

3.1 Participants

Twenty-six students from the RNNC in Den Helder, The Netherlands, were recruited to participate in this experiment, including seven females. The median age was 22 years, with a minimum of 19 and a maximum of 41 years. Due to participant dropouts (caused by deployment, courses, etc.), two teams consisted of only two participants, and one session had only one team. Only participants with a complete data set, consisting of electrocardiogram (ECG) signals; questionnaires; and video data, were included in the analyses. The final data set consisted of 10 participants; two females and eight males of whom eight had between 0 and 2 years of operational

service and two had over 2 years of operational service. The participants signed a consent form, and the study was approved by the ethical committee of Delft University of Technology, and the ethical committee of TNO.

3.2 Materials

Two static bridge simulators from the RNNC were used: the primary simulator simulated the 'Hr. Ms. Tromp' frigate (Fig. 2), and the secondary simulator simulated the 'Hr. Ms. Van Amstel' frigate. These simulators consisted of a replica of the ships bridge and virtual surroundings, such as a moving horizon that gave the perception of ship movement. To control the ships, communication was necessary between the crew on the bridge (the participants), the superiors ashore (trainers), the crew on deck (trainers), and other ships (participants and trainers). In both simulators, at least two trainers were present during the scenarios.

3.3 Measurement of variables

Seven variables were measured that appear in the COPE model as appeared in Fig. 1: (1) *events*, (2) *objective task demand*, (3) *appraisal*, (4) *subjective task demand*, (5) *emotional state/arousal*, (6) *performance*, and (7) *errors*. For every event that occurred, these variables were measured. Since there were 21 identifiable events, it was not preferred to use long questionnaires since interruptions of complex tasks lower their performance (Speier et al. 1999). For coping strategy, no short questionnaire was found so a long questionnaire was used that measured general coping and not task-specific coping. This questionnaire was filled in once. Therefore, the coping strategy measures were not used in the analyses. The different measurements are explained in the next subsections.

3.3.1 External world: stressful events

A stressful, realistic scenario was written especially for this experiment by the simulator trainers of the RNNC. In Table 1, the episodes, goals, and actions of the tasks as suggested by Ozel (2001) are described. Five main episodes were identified: (1) shadowing the smuggling ship, (2) avoiding other vessels (this goal stays a goal during the whole experiment), (3) preparing for boarding, (4) execute boarding, and (5) reacting to and execute a search and rescue (SAR). Within these main episodes, different actions can be identified as indicated in Table 1 by the letters 'a' through 'g'.

The scenario took place in the North Sea, which is familiar territory for the participants. The scenario started with two navy warships shadowing a ship that was suspected of smuggling refugees. This ship discovered that it was being followed, which means they were likely to 'destroy evidence'. In other words: throwing the refugees overboard. The participants needed to board the smuggling ship. Before the ship could be boarded, several actions needed to be taken. When the boarding was being executed, a Mayday call came in on the radio. The two Navy ships needed to decide to follow the distress call and transfer the boarding operation to another ship. When the search-and-rescue (SAR) was being executed, several actions needed to be taken. Depending on previous decisions and speed of the actions, some of the tasks could not be performed. All teams played the scenario for approximately 130 min.

3.3.2 External world: objective task demand

Several questionnaires were available for measuring task demand. A reliable, fast, and easy scale is the Overall Workload questionnaire (Hill et al. 1992). This questionnaire consists of one scale, ranging from 0 to 100. A similar single-scale questionnaire was used in this study to measure task demand assessed by the trainers. They filled in the 10-point task demand scale for novice students (0–2 years of experience) and more experienced students (more than 2 years of service). Although the measurement itself is 'subjective', the trainers rated the events as external and objective experts (i.e. not participating in the stressful situation) from the trainees' point of view. It was therefore used as measure for objective task demand as described in Sect. 2.

Fig. 2 Bridge simulator based on the 'Van Tromp' ship, seen from two angles and the trainer control room

Table 1 Actions that need to be executed in different stages of the scenario

Episode	Time in scenario	Stressful events: actions for episode goal
1. Shadowing target ship	Start to ±25 min	(a) Start of the training
		(b) Reacting when shadowing is discovered
2. Avoiding other vessels in the dark	During entire scenario	
3. Preparing to board target ship	±25 to ±90 min	(a) Deciding what team does what
		(b) Positioning of the ships
4. Executing combined boarding	±35 to ±90 min	(a) Hailing of the target ship
		(b) Positioning the target vessel
		(c) Directing the crew
		(d) Mutual communication
		(e) Reacting on incoming Mayday
5. Executing search and rescue	±90 to end	(a) Transfer target ship to arriving coastguard
		(b) Launch helicopter
		(c) Gearing up against traffic flow
		(d) Navigate between sandbars
		(e) Searching for 'man-over-board'
		(f) Deploying the medic
		(g) Carrying away injured

3.3.3 Cognition: appraisal

For every event in the scenario (Table 1), the participants filled in an appraisal questionnaire. One scale running from (1) challenge to (10) threat was filled in for every event in the scenario. With this scale, the appraisal could not be filled in as 0 but was always biased towards either challenge or threat. The scores were separated into two variables: challenge and threat. The challenge variable was created out of the scores from 1 to 5 correspond to 'very challenging' (1) and 'little challenging' (5). A threat variable was created out of the scores from 6 to 10 where 6 corresponds to 'little threatening' and 10 corresponds to 'very threatening'. The appraisal scores 5, 4, 3, 2, and 1 were reversed to challenge scores 1, 2, 3, 3, and 5. The appraisal scores 6, 7, 8, 9, and 10 were converted to the threat scores of 1, 2, 3, 4, and 5. In this manner, the two appraisal variables could be compared.

3.3.4 Cognition: subjective task demand

The subjective task demand was measured with the same questionnaire as the objective task demand. A single scale, ranging from 1 'not at all demanding' to 10 'very demanding', was filled in by the participants scoring their own (subjective) task demand.

3.3.5 Cognition: emotional state: arousal

To measure the participant's arousal levels during the experiment without having the participants fill in a questionnaire, six mobi8 systems from TMSi (Enschede, The Netherlands) were used. These devices measure electrocardiographs (ECG), which can be translated into heart rate and heart rate variability. Each mobi8 has three sensors: one sensor was placed on the right collarbone, another sensor under the left ribs, and a ground sensor was placed on the right side, as shown in Fig. 3. To ensure that participants could walk around freely, they carried the mobi8 devices in a suitable case.

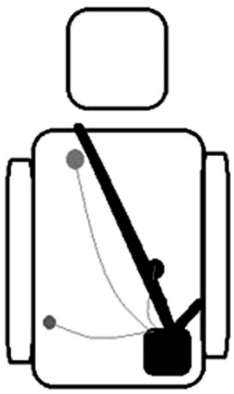

Fig. 3 Sensor placement for ECG

3.3.6 Action: performance

At the end of the experiment, the 'performance' was assessed by the trainers. All events from the session were rated on a 10-point scale for every participant. At least two trainers scored each participant, in order to create an averaged performance score. Z-scores were calculated for the performance rates, to extinguish possible trainer biases.

Table 2 Trainer comments can help in identifying the error category

Category and description	Example
Communication: Participants forget to communicate information to other participants. This is a crucial point in co-operation	The participants want to execute a boarding soon, and they are informing the crew Trainer: 'You should not yet tell them about the boarding if it is not confirmed by the commander'
Planning: When relevant information enters the bridge, it can be used to make a plan for further actions. Often, participants have the information but have not made a plan yet	The participants started a particular engine of the ship, which cannot run for longer than 15 min Trainer: 'What are you going to do with these engines? They are going to break down soon'
Speed: Speed is of major essence in this scenario. Plans need to be made fast, and actions need to be executed fast. The decision-making often takes too much time	Between the ship and the Mayday location are sandbanks. The students want to go around them Trainer: 'Why do you want to go around them? Going between them is much faster'
Task allocation: Three people are on the bridge at all times (in this setting). They all have their own task, but when needed, task can be allocated differently to relieve one person of too much tasks. This is often forgotten	One student is only focusing on reading the map Trainer: 'You should alternate between your tasks more'

Z-scores for a single participant's performance rate were calculated with the mean and standard deviation of all the performance scores from all participants.

3.3.7 Action: error

Two Sony HDR-CX300E cameras, a Sony handy cam DCR-SR55 and a Panasonic HDC-RM300 camera, were used to record the activities in the simulators. Two cameras were placed in each simulator. The video data were used to define what situation and action occurred every minute. These videos were used to observe the trainer comments that could be used to determine whether, when, and what kind of errors occurred.

Within the video data, some errors were clearly identifiable. These errors were a direct result from faulty actions (Rasmussen 1982). For example, in one team, the members were all focusing on their own task which made them forget to keep track of the radar and look outside. They did not notice a buoy in front of the ship and navigated over the buoy. Only relative few of these kinds of errors happened during the experiment. Other errors were not directly visible, but the actions taken by the participants would not meet the planned goal. These actions would only unfold into an error, after a substantial amount of time had passed (Rasmussen 1982). To identify these unfitting actions, the comments from the trainers were analysed. An example: the team members forgot to communicate their plan to the crew of the ship. If the crew does not prepare for action, the action cannot be performed when the participants want it to be performed. These errors, or more precisely, *tendencies to err*, were identified based on the comments and suggestions made by the trainers. Comments were categorized into five groups, which corresponded to the groups of cognitive issues indicated by Dowell and Hoc (1995): communication; planning; speed; task allocation; and 'other'. For every category, an example is given in Table 2.

3.4 Procedure

Five experimental sessions were performed. In an experimental session, the scenario was played with six participants divided over two teams and simulators. The scenario lasted about 2 h, with a 15-min break halfway the scenario. Each team had a participant fulfilling the role of an *officer of the watch*, a *navigation officer* and a *steersman*.

Before running the scenario, the participants gathered in a classroom where they received a briefing about the scenario and the general aim of the study from the trainers. The participants were assigned to teams and divided the roles within the teams. After this, a questionnaire was filled in, in which general information about the participants was asked: e.g. years of service; experience in the simulators; and some general health questions, e.g. do you smoke, drink alcohol, or caffeine. Next, the mobi8 systems were explained and connected to the participants. After the briefing, the participants went to the simulators where video cameras were turned on.

At the moment the simulators were started, the participants turned on the mobi8 systems that started recording ECG. The first half of the scenario was played, followed by a 15-min break, in which the participants answered the

appraisal and task demand questionnaire for every action they encountered in the first half of the scenario. The scenario was then continued. Due to differences in the decisions and actions taken by the different teams, not all sessions lasted the same amount of time. After approximately 2 h, the scenario was ended by the trainers, and the second appraisal and task demand questionnaires were filled in about the events in the second half of the scenario.

The participants returned to the classroom where they took off the mobi8 sensors and were debriefed by the trainers. After the debriefing, the participants left and the trainers filled in the performance questionnaires, rating the actions of every participant. Although the basics of the scenarios were the same during every session, decisions made by the participants led to small differences in the storyline and the order of the events.

4 Results

The results section is divided into two parts. The first part focuses on the variables of the COPE model. The second part focuses on creating a predictive model out of the data set. Before the data could be analysed, the raw measurements were first transformed into a data set ready for analysis.

4.1 Data preparation

The ECG data, as collected with the mobi8 from TMSi, were converted into heart rate (HR) and heart rate variability (HRV) per minute, using Matlab R2011a (The Mathworks). The signal measured in mV was first passed through a high-pass (0.5 Hz) and a low-pass (40 Hz) filter. After filtering, a peak-detection function was applied to the ECG signal. A minimum value had to be set in order to only detect the R-tops of the heart beat. Counting the number of R-tops per minute resulted in the HR value per minute.

Nine outliers in the HR data, defined as values larger than three times the inter-quartile range, were removed from the data set, as they probably occurred because the heart rate measurement devices had stopped, or were momentarily turned off. The HRV was calculated by the root mean squared successive differences (RMSSD) method. This method squares the average of the differences between two consecutive R-tops and was calculated for every minute.

A reliability analysis was conducted across the participants to examine similarity between participants' responses to their subjective task demand and appraisal. Table 3 shows the Cronbach's alpha values for both variables for the 26 participants and the group of $n = 10$ from the final data set. Alpha values range from 0.75 to 0.99; it seems that there was a strong correlation between the participants' appraisal and subjective task demand.

Table 3 Cronbach's alpha for appraisal and subjective task demand scores between participants and subjective task demands scores between trainers

	Cronbach's alpha	
	26 pp	10 pp
Appraisal	0.92	0.92
Subjective task demand	0.99	0.75

With the help of video data, it was determined which action (from Table 1) was executed at which time by each participant. The comments from trainers were used to determine whether errors were (almost) made by the participants. For every action, data about the appraisal, task demand, and performance were collected by means of the questionnaires described in the method section. Knowing what actions were executed every minute allowed us to calculate the appraisal, task demand, and performance per minute. If multiple tasks were performed in one particular minute, the associated appraisal and task demand scores were summed. For performance, scores were normalized and averaged per minute for all the tasks performed. Since the sessions all lasted over 2 h, around 130 data points per participant were collected. As an example, a small part of the data set is displayed in Table 4.

Besides the minute-by-minute data, six extra lag variables were created for HR, HRV, threat, challenge, objective and subjective task demand, and the errors and performance variables. These lag variables were created by taking the average value over a window of the previous 5 min. Using lag variables might result in better predictions if the effects of stress are delayed or take more time to appear than 1 min. For the error variable, the lag-variable would be a '1' if the previous 5 min would contain a '1'.

The trainer comments were coded by three independent coders into five categories (Table 2). Coder 1, the experiment leader, coded the comments into the five categories and made a description of the categories. These were explained to coder 2 and 3. The first round of codes was examined, and the non-matching codes were discussed. Then, coders 2 and 3 coded the comments a second and a third time. As can be seen in Table 5, coder 2 fully agreed with the coding of coder 1 while coder 3 had some disagreements. Table 5 shows the Cohen's kappa for inter-rater agreement. The inter-rater agreement ranges between 0.72 and 1, except for the 'other' category that had the lowest inter-rater correlation of 0.46. This category was therefore left out of the analyses.

Table 4 A small part of the complete data set

pp	Time	HR	HRV	Appraisal		Task demand			
				Threat	Challenge	Objective	Subjective	Performance	Error
2	1	104.32	0.58	0	1	4.50	5	−0.79	0
2	2	97.75	0.61	2	1	8.50	13	−0.20	0
2	3	98.03	0.61	2	0	4.00	8	0.39	0
2	4	97.07	0.61	0	1	4.50	5	−0.79	0
2	5	99.73	0.60	0	6	5.67	0	0.00	0
2	6	101.65	0.59	0	6	5.67	0	0.00	0
2	7	97.72	0.61	2	6	9.67	8	0.19	0
2	8	104.82	0.57	2	0	4.00	8	0.39	1
2	9	101.16	0.59	0	6	5.67	0	0.00	0
2	10	107.49	0.56	0	7	10.17	5	−0.40	0

The columns indicate; participant, minute, heart rate, heart rate variability, appraisal (threat and challenge) task demand (objective and subjective) normalized performance and the error status (0 = No, 1 = Yes)

Table 5 Cohen's kappa for the inter-rater correlations between 3 raters and 5 categories

Category	Coder 1 Coder 2	Coder 1 Coder 3
Communication	1.00	0.77
Planning	1.00	0.72
Speed	1.00	0.80
Task allocation	1.00	0.76
Others	1.00	0.46

4.2 COPE model exploration

The first step into the exploration of the COPE model was to examine the different variables and therewith testing the first hypothesis. Table 6 shows the sample size, minimum and maximum score, mean and standard deviation for each variable in the data set. There are less data points for the lag variables than for the non-lag variables, because the lag variables were calculated starting at the sixth minute of the session. After removing the heart rate outliers, the lowest heart rate recorded is 45.48 beats per minute, and the highest is 116.82 beats per minute. It is interesting to note that the mean of the normalized performance lies below 0. The error scores are either one or zero. The mean scores for all the error variables are close to zero, which illustrates an underrepresentation in the error data, which will be discussed later in this paper.

Next, the correlations between the different variables were examined. To control for between participants variance, correlations of the variables were first calculated per participant and then averaged. The average amount of data points per participant is 116, which gives a df of 114. The critical correlation value for $df = 114$, and $\alpha = 0.05$ is $r_c = 0.179$. Table 7 shows all the correlations. Those bold faced are significant, and those bold faced and highlighted are significant correlations between different variables.

Table 7 shows a negative correlation between heart rate and heart rate variability; higher HR correlates to lower HRV and vice versa. These relations can be seen in both the 5-min lag measure and the 1-min measure. Among the regular variables, six significant correlations were found. Challenge and threat appraisals show a negative correlation as expected since appraisal was measured on a single scale ranging from challenge to threat. Objective and subjective task demand correlated positively, indicating that participant and trainer perception corresponded to each other. Likewise, a positive correlation was found between task demands and both threats and challenge appraisals. This suggests that low task demand situations were not likely to be appraised as a threat or a challenge, while highly demanding situations were.

The correlations between the lag variables show similar patterns, with two exceptions: a challenge appraisal was no longer found to correlate with subjective task demand, but was found to correlate positively with hearth rate and negatively with heart rate variability. In other words, this result supports the COPE model's link between arousal and challenge appraisal. Interestingly, no direct correlations were found between variables from the model and the minute-by-minute performance and errors (Table 7). Still, on a 5-min window, the lag variables show that challenge appraisal was reversely correlated with errors.

4.3 Predictive models

Four generalized linear mixed model (GLMM) analyses were conducted to analyse the relation between the COPE model variables and the observed performance and cognitive errors. These analyses tested the second general hypothesis of this study. Performance and errors were modelled as dependent variables, using a linear model and a binary logistic regression model, respectively. The fixed factors consisted of the independent variables HR, HRV, threat, challenge, objective task demand and subjective

Table 6 Descriptive statistics of the model's variables and the lag variables

	N	Minimum	Maximum	Mean	SD
Emotional state (arousal)					
Heart rate	1,168	45.48	116.82	80.96	12.56
Heart rate variability	1,168	0.51	1.38	0.76	0.14
Appraisal					
Threat	1,168	0	8	0.68	1.43
Challenge	1,168	0	20	4.76	3.75
Task demand					
Objective	1,168	0	24.33	8.31	4.54
Subjective	1,168	0	26.00	7.20	5.55
Actions					
Performance	1,168	−3.15	1.57	−0.45	1.06
Errors	1,168	0	1	0.09	0.28
Communication	1,168	0	1	0.04	0.19
Planning	1,168	0	1	0.04	0.20
Speed	1,168	0	1	0.01	0.12
Task allocation	1,168	0	1	0.01	0.11
Other	1,168	0	1	0.02	0.14
Lag variables					
Heart rate	1,119	50.50	109.59	81.01	12.06
Heart rate variability	1,119	0.55	1.22	0.76	0.14
Appraisal threat	1,119	0.00	6.20	0.69	1.29
Appraisal challenge	1,119	0.00	15.98	4.71	3.32
Objective task demand	1,119	1.80	18.93	8.29	3.48
Subjective task demand	1,119	0.00	19.60	7.20	4.64
Performance	875	−3.15	1.57	−0.43	0.96
Error	1,103	0	1	0.37	0.48

task demand, and their the lag variables. 'Participant' was included as a random factor, thereby including a random intercept for each participant. The variance component type was used for random effect covariance type.

4.3.1 Performance

A GLMM shows that the fixed factors can explain the performance per minute [$F(6,1.161) = 8.60$, $p < 0.01$] with a correlation of $r = 0.77$ between observed and predicted performance. The individual variance differed significantly from the standard intercept ($\text{var}_{\text{intercept}} = 0.718$, Std Err = 0.35, $Z = 2.08$, $p = 0.037$), indicating that on average the participants differed in their performance variance among each other. Examining the coefficients in Table 8 shows that an increase in threat or challenge appraisal coincided with significant increase in the performance. The analysis shows an opposite effect for objective task demand. An increase in this factor coincided with significant decrease in performance. Including the lag variables in the GLMM analysis resulted again in a model with explaining ability [$F(12,1.106) = 5.99$, $p < 0.01$] with a correlation of $r = 0.79$ between predicted and objective performance. Also this model shows a significant random intercept for individual participants ($\text{var}_{\text{Intercept}} = 0.723$, Std Err = 0.35, $Z = 2.06$, $p = 0.039$). In addition to factors already found in the previous model, the extended model also revealed that an increase lagged threat appraisal of the last 5 min coincided with reduction in performance (Table 9).

4.3.2 Predictive error models

The GLMM analysis revealed a significant binary logistic model for the error variable, $F(6,1.161) = 5.57$, $p < 0.01$. On average, the model predicted 91.2 % of the error status correctly, with 100 % correct predictions for 'no error', and 0 % correct predictions for 'error'. The model found no significant ($\text{var}_{\text{intercept}} = 0.195$, Std. Err = 0.198, $Z = .984$, $p = 0.33$) difference between the participants with regard to making an error. Table 10 shows that an increase in challenge appraisal coincided with an increased chance of making an error. Extending the model with lag variables resulted again in a significant model [$F(12,1.106) = 4.29$, $p < 0.01$], however, without any significant coefficient (all $p > 0.05$).

Table 7 Correlations between all the variables

	Regular variables						Lag variables						
	HR	HRV	Appraisal threat	Appraisal challenge	Objective task demand	Subjective task demand	HR	HRV	Appraisal threat	Appraisal challenge	Objective task demand	Subjective task demand	Performance
Regular variables													
HR	1.00												
HRV	**−0.99**	1.00											
Appraisal threat	0.03	−0.03	1.00										
Appraisal challenge	0.14	−0.13	**−0.23**	1.00									
Objective task demand	0.09	−0.08	**0.40**	**0.63**	1.00								
Subjective task demand	−0.01	0.02	**0.47**	**0.23**	**0.75**	1.00							
Lag variables													
HR	**0.44**	**−0.44**	0.09	**0.18**	0.10	−0.03	1.00						
HRV	**−0.48**	**0.49**	−0.09	**−0.19**	−0.09	0.06	**−0.96**	1.00					
Appraisal threat	0.03	−0.04	**0.41**	−0.17	0.07	0.14	0.07	−0.09	1.00				
Appraisal challenge	0.12	−0.12	**−0.21**	**0.45**	0.17	−0.10	**0.23**	**−0.22**	**−0.27**	1.00			
Objective task demand	0.08	−0.07	0.03	**0.21**	**0.37**	**0.22**	0.14	−0.13	**0.32**	**0.53**	1.00		
Subjective task demand	−0.02	0.04	0.13	−0.04	**0.24**	**0.46**	0.01	0.03	**0.42**	0.06	**0.71**	1.00	
Performance	0.03	−0.03	0.02	0.08	0.03	0.05	0.03	−0.05	−0.10	−0.01	−0.03	−0.05	1.00
Error	0.02	0.00	−0.03	−0.12	−0.10	−0.05	−0.01	0.01	0.08	−0.03	0.03	0.06	−0.05
Lag													
Performance							−0.03	0.03	0.07	0.09	0.12	0.17	
Lag													
Error							0.00	0.04	0.08	**−0.24**	−0.07	0.06	−0.00

Calculated by averaging the correlations between variables for every participant

Bold-faced values are significant at $\alpha = 0.05$

$n = 116$, $df = 114$, $r_c = 0.18$

Table 8 Results of GLMM analysis on performance without lag variables

	df1	*df2*	*F*	Sig	Coefficient	Std error	*t*	Sig.	Lower	Upper
Corrected model	6	1.161	8.60	**<0.01**						
HR	1	1.161	0.18	0.68	0.00	0.01	−0.42	0.68	−0.02	0.02
HRV	1	1.161	0.78	0.38	−0.73	0.83	−0.88	0.38	−2.36	0.9
Appraisal threat	1	1.161	20.46	**<0.01**	0.12	0.03	4.52	**<0.01**	0.07	0.18
Appraisal challenge	1	1.161	33.67	**<0.01**	0.07	0.01	5.8	**<0.01**	0.05	0.09
Objective task demand	1	1.161	22.99	**<0.01**	−0.06	0.01	−4.8	**<0.01**	−0.08	−0.04
Subjective task demand	1	1.161	0.62	0.43	0.01	0.01	0.79	0.43	−0.01	0.03
Intercept					0.42	1.45	0.29	0.77	−2.42	3.26

Bold *p* values are significant

Table 9 Results of GLMM analysis on performance with lag variables

	df1	*df2*	*F*	Sig	Coefficient	Std	*t*	Sig	Lower	Upper
Corrected model	12	1.106	5.99	**<0.01**						
HR	1	1.106	0.58	0.45	−0.01	0.01	−0.76	0.45	−0.03	0.01
HRV	1	1.106	0.72	0.407	−0.80	0.94	−0.85	0.40	−2.64	1.05
Appraisal threat	1	1.106	34.64	**<0.01**	0.19	0.03	5.89	**<0.01**	0.13	0.26
Appraisal challenge	1	1.106	21.14	**<0.01**	0.07	0.01	4.60	**<0.01**	0.04	0.10
Objective task demand	1	1.106	18.65	**<0.01**	−0.06	0.01	−4.32	**<0.01**	−0.09	−0.03
Subjective task demand	1	1.106	0.37	0.55	0.01	0.01	0.61	0.55	−0.02	0.03
Lag_HR	1	1.106	0.07	0.80	0.00	0.01	−0.26	0.80	−0.03	0.02
Lag_HRV	1	1.106	0.60	0.44	−0.86	1.11	−0.77	0.44	−3.03	1.32
Lag_appraisal threat	1	1.106	15.48	**<0.01**	−0.17	0.04	−3.93	**<0.01**	−0.25	−0.08
Lag_appraisal challenge	1	1.106	0.16	0.69	−0.01	0.02	−0.40	0.69	−0.04	0.03
Lag_objective task demand	1	1.106	0.00	0.96	0.00	0.02	0.05	0.96	−0.04	0.04
Lag_subjective task demand	1	1.106	0.18	0.67	0.01	0.01	0.42	0.67	−0.02	0.03
Intercept					1.78	1.83	0.97	0.33	−1.81	5.37

Bold *p* values are significant

Table 10 Multilevel linear regression for error prediction without lag variables

	df1	*df2*	*F*	Sig	Coefficient	Std	*t*	Sig	Exp coefficient	Confidence interval for exp (coefficient)	
										Lower	Upper
Corrected model	6	1,161	5.57	**<0.01**							
HR	1	1,161	1.83	0.18	0.11	0.08	1.35	0.18	1.12	0.95	1.32
HRV	1	1,161	3.24	0.07	17.04	9.46	1.80	0.07	25.05×10^{6}	0.22	2.88×10^{15}
Appraisal threat	1	1,161	2.45	0.12	0.21	0.13	1.57	0.12	1.23	0.95	1.59
Appraisal challenge	1	1,161	5.64	0.02	0.15	0.06	2.37	**0.02**	1.16	1.03	1.31
Objective task demand	1	1,161	3.08	0.08	0.10	0.05	1.76	0.08	1.1	0.99	1.22
Subjective task demand	1	1,161	1.60	0.21	−0.06	0.05	−1.27	0.21	0.94	0.86	1.03
Intercept					−20.51	13.79	−1.49	0.14	0.00	0.00	700.79

Bold *p* values are significant

The analysis of errors led to two important observations: (1) as only 91.2 % (1,065/1,168) of intervals included no error, the prediction was strongly biased towards no error prediction and (2) no individual difference between participants was found. Such an error prediction model would not be useful in a training setting. Instead, in such a setting, it would be acceptable to have some level of false alarms, if it would increase the number of correct predicted errors, i.e. hits. Therefore, analyses were also conducted that corrected for the bias

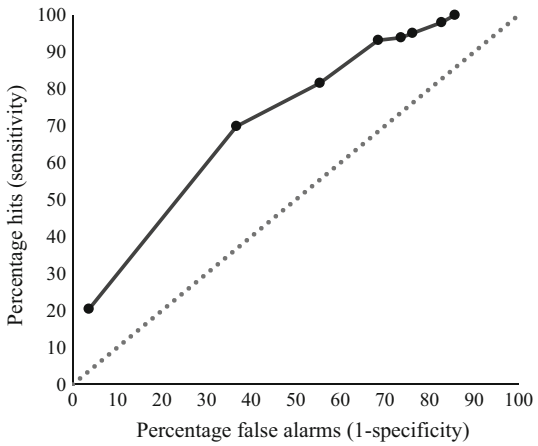

Fig. 4 ROC curve consisting of logistic regressions for the error variable with different weighted cases

towards no error and no longer included participants as a random intercept, i.e. a normal logistic regression was deemed sufficient.

The underrepresentation of errors in the data set was corrected by giving weights to the different cases. A receiver operation characteristic curve (ROC curve) was used to determine the proportion for the case weighting that gives the most optimal logistic regression results. Figure 4 shows the ROC curve with the false alarm rate on the *x*-axis and the hit rate on the *y*-axis. Two methods for determining the optimal weighting were used, namely the closest point to the ideal situation of 100 % hits and 0 % false alarm (d^2), and the maximum sum of sensitivity (*Sn*) plus specificity (*Sp*) (Kumar and Indrayan 2011). Applying these methods, a weight ratio of 90:10 was determined for error versus no error, as this ratio resulted in a logistic regression with the highest sum of specificity plus sensitivity (1.33) and the shortest distance from the ideal left upper corner of ROC (distance = 0.23). Applying this weighing ratio led to significant logistic regression model [$\chi^2(6, n = 1{,}168) = 3{,}761.26, p < 0.01$] that included an intercept and the other COPE model variables. Whereas the logistic regression model with only an intercept had a correct prediction rate of 53.5 %, adding the variables improved this to 66.4 %, with an Cox and Snell's $R^2 = 0.17$. Adding the lag variables also created a significant model [$\chi^2(12, n = 1{,}168) = 4{,}567.445, p < 0.01$] with a correct prediction rate of 68.3 % and a Cox and Snell's $R^2 = 0.21$. This model had a correct prediction of 52.3 % when only the intercept was used. As Table 11 shows, all the coefficient in the model are significant (all $p < 0.05$).

This same procedure was also used to conduct logistic regression analysis on the specific type of errors, i.e. communication, planning, speed, and task allocation. Table 12 shows the different weighting ratios used for each error category. The correct prediction ranged from 66.4 % for planning errors to 91.5 % for task allocation errors. All logistic regression models were significant ($p < 0.05$) with Cox and Snell's R^2 ranging from 0.19 to 0.55.

4.4 Cross-validations

To test the generalizability of the performance model, a cross-validation was conducted (Refaeilzadeh et al. 2009). This means that the data set was divided into two sets: one to train the model and one to validate the model. The leave-one-out cross-validation, a specific form of *k*-fold cross-validation, was applied. Here, the data set was divided into ten parts. Data from nine participants were used as the training part to create the regression model, i.e. determine the coefficients. This would lead to formulas with a general form:

$$\begin{aligned}\text{Predicted performance} &= \text{intercept} + (b * \text{Heart Rate}) \\ &+ (b * \text{Heart Rate Variability}) + (b * \text{Threat}) \\ &+ (b * \text{Challenge}) + (b * \text{Objective Task Demand}) \\ &+ (b * \text{Subjective Task Demand})\end{aligned}$$

Data from the participant that was left out were used as the validation part of the model by entering the actual values of the predictors, included the lag variables, and calculating the predicted performance. Every participant was used once as the validation part, which created predictive performance values for all the participants.

The predicted performance values from a GLMM (including lag variables) without random factors, also known as a linear regression, correlated with observed performance values ($r = 0.56$). A cross-validation for this model still showed a significant correlation, although reduced [$r(1{,}168) = 0.17, p < 0.01$].

A similar procedure conducted for the weighted logistic regression model on the cognitive error, in general, where the total logistic regression model (including lag variables) correlated with the observed errors with an $r = 0.23$, the cross-validation model lowered this correlation with the observed errors to $r(1{,}165) = 0.13$, $p < 0.01$. This cross-validation model for the errors had a correct prediction of 67.3 %, which is close to the 68.3 % correct prediction for the model based on total sample.

5 Discussion and conclusion

The first hypothesis of this study states that there are relationships between the variables of the COPE model. As the correlation table shows, correlations exist between the variables. Only the physiological variables of heart rate and heart rate variability do not seem to correlate to the other cognitive or performance variables.

The second hypothesis was also confirmed. Models were created that use situational and cognitive variables to

Table 11 Results of weighted logistic regression for the error variable including lag variables

B	S.E.	Wald	*df*	Sig.	Exp(β)	
HR	−0.04	0.02	5.32	1	0.02	0.964
HRV	−7.61	1.76	18.66	1	<0.01	4.95×10^{-4}
Appraisal threat	−0.36	0.03	208.30	1	<0.01	0.696
Appraisal challenge	−0.14	0.01	139.39	1	<0.01	0.868
Objective task demand	−0.07	0.01	36.99	1	<0.01	0.937
Subjective task demand	0.03	0.01	13.81	1	<0.01	1.034
Lag_HR	−0.21	0.02	119.41	1	<0.01	0.813
Lag_HRV	−21.29	2.07	106.22	1	<0.01	5.67×10^{-10}
Lag_appraisal threat	0.33	0.03	144.09	1	<0.01	1.39
Lag_appraisal challenge	0.03	0.01	6.93	1	0.01	1.035
Lag_objective task demand	0.08	0.01	43.09	1	<0.01	1.079
Lag_subjective task demand	0.02	0.01	4.72	1	0.03	1.021
Intercept	41.12	2.43	287.36	1	<0.01	7.18×10^{17}

Table 12 Logistic regressions for the four error categories with lag variables

	Optimal case weight	d^2	Sn + Sp	Model	Correct predictions for intercept model and for intercept + variables model (%)	Cox and Snell's R^2
Communication	98:02	0.141	1.48	χ^2 (12, n = 1,119) = 2,288.835, $p < 0.05$	51.3–74.1	0.34
Planning	98:02	0.266	1.308	χ^2 (12, n = 1,119) = 1,180.384, $p < 0.05$	53.9–66.4	0.19
Speed	99:01	0.194	1.385	χ^2 (12, n = 1,119) = 860.390, $p < 0.05$	56.3–69.9	0.24
Task allocation	99:01	0.017	1.822	χ^2 (12, n = 1,119) = 2,219.745, $p < 0.05$	64.3–91.5	0.55

predict performance and errors. Tables 8, 9 and 11 show the contribution of the variables to the different outcome variables. Figure 5 shows how performance and errors can be predicted out of the COPE variables. The analyses used in this study showed how much of the variance in the performance and error variables was accounted for by the COPE model's variables. The significant predictions found in the analyses are presented as arrows in Fig. 5. Performance rated by experts can be predicted out of the threat, challenge, and objective task demand variables (solid lines), but not out of the physiological measures of arousal. Participants walked around in the simulator, and this might have been a distorting factor in the measurements of arousal. The strong correlation that was found between HR and HRV but not between HR or HRV and the other variables might show a ceiling effect.

Errors, on the other hand, could be predicted out of the physiological measures, which might indicate that the method used for scoring 'expert-rated-performance' might have been un-synchronized with the ECG measures. The errors were extracted every minute from the videos and are therefore better synchronized with the ECG measures that were also measured per minute. The performance scores were measured with questionnaires that listed the executed tasks in the same way as the questionnaires for appraisal and task demand did. Future studies should look into the combination of different measurement systems and how to improve synchronization between these different measurements.

Errors can be predicted out of all variables (dotted lines). The ability to predict errors varies between error categories, with planning errors having the lowest and task allocation errors having the highest correct prediction rates. Furthermore, the cross-validation analysis showed the possibility of making a significant prediction for a new data set, suggesting the generalisation of these prediction models. Other studies have done similar research, but within different context and with different methods. As a first example, Neerincx et al. (2009) created a naïve Bayesian network to predict performance of naval operators. The COPE model includes Neerincx' model, addressing more factors and distinguishing several error types, and can therefore be used for training purposes. A second example is the Structural Equation Model of Kylesten (2013) that describes dynamic decisions-making on operative levels. Kylesten (2013) also used a descriptive model to describe dynamic decision-making and fitted data to this model. In contrast to the COPE model, this model did not include an objective measure from an instructor, and no physiological measures were used.

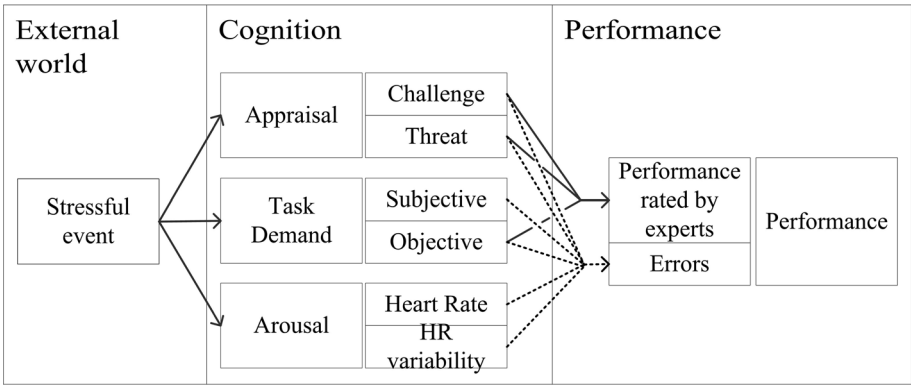

Fig. 5 COPE model with indications of validated correlations, and validated predictive values

This study has a number of limitations that should be noted. Although observation data were collected from 26 participants, only data of 10 participants were included in the analysis, giving this study a small sample size regarding the number of participants involved. When it comes to the amount of 1-min observations, this study had a relatively large sample ($n = 1{,}168$). This sample ratio seems appropriate as the focus of the work was not to examine performance and cognitive errors between individuals, but between different stressful situations within subjects. Cross-validation analyses showed a reduction in prediction accuracy compared to the GLMM models, but the predictions still correlated significantly with the observation data. This supports the prediction models' ability to generalize outside the sample of individuals included in this study. Another limitation was that the data were collected within teams, and therefore, individual observations might not be completely independent. Future studies that include more individuals might consider to include different teams as a random factor in the analysis. Future studies might also consider the effect of different individual characteristics, as this study found that performance prediction differed between the participants. For the arousal measurements, other physical indicators, such as galvanic skin response, might be more suitable for a setting in which physical movement is inevitable.

There are several ways to increase the prediction accuracy of the models. First, broadening the 1-min interval prediction window, for example, to 5 min might lead to higher accuracy in the predictions. Compared to the 1-min performance and error variables, correlations between the 5-min performance and error variables and the other variables are slightly stronger, with one significant correlation. It might be easier to predict over a longer period of time, but for a fast-paced stressful training scenario, it might not be appropriate to deliver feedback for a 5-min period; hence, this paper mainly had a minute-by-minute focus.

Second, in this study, cognitive errors were defined as an intervention or comment by the trainers. When exactly a trainer decides to intervene or make a comment, might vary and the predictions per minute are likely to be error prone. Therefore, when giving minute-by-minute error feedback in a training situation, giving error likelihood feedback might be more appropriate than a simple yes or no error type of feedback.

A third way to improve the models prediction accuracy might be to add information about the participants coping strategies. As can be seen in Figs. 1 and 5, coping strategy is an intervening variable between the other cognitive variables and the actions of the individual. According to the COPE model, the data used to predict the errors and performance were all indirect factors and therefore less able to provide information for accurate prediction.

Besides the support found for the COPE model, the second contribution of this paper is the demonstration of creating a model for minute-by-minute predictions of performance and cognitive errors in a virtual stressful situation. When using such a model, the necessary information needs to be available per minute, in this case: the stressful environment, task demand, appraisal, and arousal. Arousal data could be obtained from physiological indicators. Assuming application of the models for the same training scenario as presented in this study, the same trainer data about the objective task demand could be used again. In an integrated environment, e.g. a virtual environment, a computer generates specific events in the training scenario, which provides the information about the stressful situation. Every event can be linked, for example, to a look-up table that holds the corresponding information about objective task demands for every event. In this study, the subjective task demand and appraisal information were obtained from students after completion of the scenario. For a minute-by-minute feedback system, this would not be suitable, since the information is needed every minute. Asking the trainees to provide this information, each time they are confronted with a new task would provide individual real-time information, but is too obtrusive and will lower the performance of the task (Speier et al. 1999) and affect their engagement or feeling of being present in such a situation (Hartanto et al. 2012) A less interruptive way

would be to use the data provided by participants in this study as a more general appraisal and subjective task demand. This last approach seems possible since high similarities were found between the participants' item responses (Table 3).

The methods suggested in this paper are in principle not limited to the training scenario used in this study. When applying it for other training scenarios, the variables related to the tasks (appraisal, task demand) need to be remeasured for every action or event occurring in that scenario. This will lead to new task coefficients that can be implemented in the created predictive models.

To conclude, the observational study and analysis presented in this paper give an overview of which variables are important when making decisions in stressful situations and present a method to predict performance and errors from these variables. With the creation of predictive models, the next step is to implement them in a feedback system for training purposes as described in the introduction. Professionals would get real-time feedback on their expected performance and the possibility of making errors, based on their current state and the state of the external world. Training decision-making under stress while receiving feedback would hopefully lead to an increase in performance and a diminishing of errors in real-live scenarios.

Acknowledgments The work presented in this paper was supported by the Dutch FES program: Brain and Cognition: Societal Innovation (project no. 056-22-010). We would like to thank the Royal Netherlands Naval College (RNNC) in Den Helder, The Netherlands, and especially the trainers at the bridge simulator for their help with running the experiment and gathering and scheduling the participants.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

Beach LR, Lipshitz R (1993) Why classical decision theory is an inappropriate standard for evaluating and aiding most human decision making. In: Klein GA, Orasanu J, Calderwood R, Zsambok CE (eds) Decision making in action: models and methods. Ablex Publishing Corporation, Norwood, pp 3–20

Bosse T, Both F, Lambalgen Rv, Treur J (2008) An agent model for a human's functional state and performance. In: International Agent Technology, Sydney. IEEE, pp 302–307

Bouchard S, Bernier F, Boivin E, Morin B, Robillard G (2012) Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. PLoS ONE 7(4):e36169

Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavioural therapy and experimental psychiatry 25(1):49–59

Cohen MS (1993) The bottom line: naturalistic decision aiding. In: Klein GA, Orasanu J, Calderwood R, Zsambok CE (eds) Decision making in action: models and methods. Ablex Publishing Corporation, Norwood, pp 3–20

Cohen I, Brinkman W-P, Neerincx MA (2012) Assembling a synthetic emotion mediator for quick decision making during acute stress. In: Proceedings of the 2012 European conference on cognitive ergonomics. ACM, Edinburgh

Delahaij R (2009) Coping under acute stress: the role of person characteristics. Kon. Broese & Peereboom, Breda

Dowell J, Hoc J-M (1995) Coordination in emergency operations and the tabletop training exercise. Le Travail Humain 58(1): 85–102

Driskell JE, Johnston JH (1998) Stress exposure training. In: Cannon-Browers JA, Salas E (eds) Making decision under stress: implications for individual and team training, vol 3., American Psychological AssociationWashington, DC, pp 191–218

Driskell JE, Johnston JH (2006) Stress exposure training. In: Cannon-Browers JA, Salas E (eds) Making decisions under stress, vol 3. American Psychological Association, Washington, DC

Forgas JP (1995) Mood and judgement: the affect infusion model (AIM). Psychol Bull 117(1):39–66

Gaillard A (2007) Stress productiviteit en gezondheid, vol 3. Holland Graphics, Amsterdam

Gohm CL, Baumann MR, Sniezek JA (2001) Personality in extreme situations: thinking (or not) under acute stress. J Res Pers 35:388–399

Haag A, Goronzy S, Schaich P, Williams J (2004) Emotion recognition using bio-sensors: first steps towards an automatic system. Affective dialogue systems, tutorial and research workshop. Kloster Irsee, Germany

Hartanto D, Kang N, Brinkman W-P, Kampmann IL, Morina N, Emmelkamp PMG, Neerincx MA (2012) Automatic mechanisms for measuring subjective unit of discomfort. Annu Rev Cybertherapy Telemed 181:192–196

Hill SG, Iavecchia HP, Byers JC, Bittner AC, Zaklad AL, Christ RE (1992) Comparison of four subjective workload rating scales. Hum Factors 34(4):429–439

Hjortskov N, Rissen D, Blangsted AK, Fallentin N, Lundberg U, Søgaard K (2004) The effect of mental stress on heart rate variability and blood pressure during computer work. Eur J Appl Physiol 92:84–89

Hollnagel E, Woods DD (2005) Joint cognitive systems: foundations of cognitive systems engineering. CRC Press, Taylor & Francis group, Boca Raton

Keinan G, Friedland N, Ben-Porath Y (1987) Decision making under stress: scanning of alternatives under physical threat. Acta Psychol 64:219–228

Kerstholt JH (1994) The effect of time pressure on decision-making behaviour in a dynamic task environment. Acta Psychol 86:89–104

Kleider HM, Parrott DJ, King TZ (2010) Shooting behaviour: How working memory and negative emotionality influence police officer shoot decisions. Appl Cogn Psychol 24:707–717

Kontogiannis T, Kossiavelou Z (1999) Stress and team performance: principles and challenges for intelligent decision aids. Saf Sci 33:103–128

Krantz G, Forsman M, Lundberg U (2004) Consistency in physiological stress responses and electromyographic activity during induced stress exposure in women and men. Integr Physiol Behav Sci 2:105–118

Kumar R, Indrayan A (2011) Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr 48:277–287

Kylesten B (2013) Dynamic decision-making on an operative level: a model including preconditions and working method. Cogn Technol Work 15:197–205

Lazarus RS (1999) Stress and emotion: a new synthesis. Springer, New York

Maule AJ, Hockey GRJ, Bdzola L (2000) Effects of time-pressure on decision-making under uncertainty: changes in affective state and information procession strategy. Acta Psychol 104:283–301

Mehrabian A (1996) Pleasure–arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr Psychol 14(4):261–292

Mendl M (1999) Performing under pressure: stress and cognitive function. Appl Anim Behav Sci 65:221–244

Mulder L (1992) Measurement and analysis methods of heart rate and respiration for use in applied environments. Biol Psychol 34(2):205–236

Neerincx MA (2003) Cognitive task load design: model, methods and examples. In: Hollnagel E (ed) Handbook of cognitive task design. Lawrence Erlbaum Associates, Mahwah, pp 283–305

Neerincx M, Kennedie S, Grootjen F, Grootjen M (2009) Modelling the cognitive task load and performance of naval operators. In: Schmorrow DD, Estabrooke IV, Grootjen M (eds) Lecture notes in artificial intelligence. Foundations of augmented cognition: neuroergonomics and operational neuroscience. Proceedings of the 5th international conference of the augmented cognition, pp 260–269

Orasanu JM, Backer P (1996) Stress and military performance. In: Driskell JE, Salas E (eds) Stress and human performance. Lawrence Erlbaum Associates Inc, Mahwas, pp 89–125

Orasanu J, Connolly T (1993) The reinvention of decision making. In: Klein GA, Orasanu J, Calderwood R, Zsambok CE (eds) Decision making in action: models and methods. Ablex Publishing Corporation, Norwood, pp 3–20

Ozel F (2001) Time pressure and stress as a factor during emergency egress. Saf Sci 38:95–107

Peeters M, Van Den Bosch K, Meyer J-JC, Neerincx MA (2014) The design and effect of automated directions during scenario-based training. Comput Educ 70:173–183

Rasmussen J (1982) Human errors: a taxonomy for describing human malfunction in industrial installations. J Occup Accid 4:311–333

Reason J (1987) Cognitive aids in process environments: prostheses or tools? Int J Man Mach Stud 27:463–470

Refaeilzadeh P, Tang L, Liu H (2009) Cross validation. In: Liu L, Ozsu MT (eds) Encyclopedia of database systems. Springer, Berlin, p 6

Sime J-A (2007) Designing emergency response training: seven ways to reduce stress. In: International conference on cognition and exploratory learning in digital age. IADIS, Algarve, PT

Speier C, Valacich JS, Vessey I (1999) The influence of task interruption on individual decision making: an information overload perspective. Decis Sci 30(2):337–360

Starcke K, Brand M (2012) Decision making under stress: a selective review. Neurosci Biobehav Rev 36:1228–1248

Thunholm P (2004) Decision-making style: Habit, style or both? Pers Individ Differ 36:931–944

Thunholm P (2008) Decision-making styles and physiological correlates of negative stress: is there a relation? Cogn Neurosci 49:213–219